Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide

Jacob Kaplan

2021-03-28

Contents

1 Preface **7**

1.1 Motivation 8

1.2 Structure of the book 8

1.3 Citing this book 9

1.4 Sources of UCR data 9

1.5 Where to find the data used in this book 11

1.6 NIBRS data 11

1.7 This book … in one paragraph 12

1.8 How to contribute to this book 12

2 Overview of the Data **14**

2.1 Crimes included in the UCR datasets 16

2.1.1 Crimes in the Offenses Known and Clearances by Arrest dataset 17

2.1.2 Crimes in the Arrests by Age, Sex, and Race dataset . 19

2.1.3 What is an index (or part 1) crime? 21

2.2 Issues common across UCR datasets 26

2.2.1 Negative numbers 28

2.2.2 Agency population value 29

2.2.3 Reporting is voluntary … so some agencies don't (or report partially) {voluntary} 30
2.2.4 Zero crimes vs no reports 31
2.3 A summary of each UCR dataset 36
2.3.1 Offenses Known and Clearances by Arrest 37
2.3.2 Arrests by Age, Sex, and Race 37
2.3.3 Law Enforcement Officers Killed and Assaulted (LEOKA) 38
2.3.4 Supplementary Homicide Reports (SHR) 38
2.3.5 Hate Crime Data 39
2.3.6 Property Stolen and Recovered (Supplement to Return A) 40
2.3.7 Arson 41
2.4 How to identify a particular agency (ORI codes) 42
2.5 The data as you get it from the FBI 43

3 Offenses Known and Clearances by Arrest 47
3.1 A brief history of the data 47
3.1.1 Changes in definitions 47
3.2 What does the data look like? 48
3.3 What variables are in the data? 48
3.3.1 Key variables 48
3.3.2 Known issues with the data 48
3.4 Final thoughts 48
3.5 Hierarchy Rule 48
3.6 Which crimes are included? 48
3.6.1 Index Crimes 49
3.6.2 The problem with using index crimes 51

3.6.3 Rape definition change 52
3.7 Actual offenses, clearances, and unfounded offenses 53
3.7.1 Actual 53
3.7.2 Total Cleared 53
3.7.3 Cleared Where All Offenders Are Under 18 54
3.7.4 Unfounded 54
3.8 Number of months reported 54

4 Arrests by Age, Sex, and Race 55
4.1 A brief history of the data 55
4.1.1 Changes in definitions 55
4.2 What does the data look like? 56
4.2.1 Raw data 56
4.3 What variables are in the data? 56
4.3.1 Key variables 56
4.4 Known issues with the data 56
4.5 Final thoughts 56

5 Law Enforcement Officers Killed and Assaulted (LEOKA) 57
5.1 Common uses of this data 58
5.2 A brief history of the data 58
5.2.1 Changes in definitions 58
5.3 What does the data look like? 58
5.4 What variables are in the data? 58
5.4.1 Key variables 58
5.5 Known issues with the data 58
5.6 Final thoughts 58

6 Supplementary Homicide Reports (SHR) 59
6.1 A brief history of the data . . . 60
6.1.1 Changes in definitions . . . 60
6.2 What does the data look like? . . . 60
6.2.1 Raw data . . . 60
6.3 What variables are in the data? . . . 60
6.3.1 Key variables . . . 60
6.4 Known issues with the data . . . 60
6.5 Final thoughts . . . 60

7 Hate Crime Data 61
7.1 A brief history of the data . . . 62
7.1.1 Changes in definitions . . . 62
7.2 What does the data look like? . . . 62
7.2.1 Raw data . . . 62
7.3 What variables are in the data? . . . 62
7.3.1 Key variables . . . 62
7.4 Known issues with the data . . . 62
7.5 Final thoughts . . . 62

8 Property Stolen and Recovered (Supplement to Return A) 63
8.1 A brief history of the data . . . 64
8.1.1 Changes in definitions . . . 64
8.2 What does the data look like? . . . 64
8.2.1 Raw data . . . 64
8.3 What variables are in the data? . . . 64
8.3.1 Key variables . . . 64
8.4 Known issues with the data . . . 64
8.5 Final thoughts . . . 64

9 Arson **65**

10 County-Level Detailed Arrest and Offense Data **66**

10.1 A brief history of the data 66

10.1.1 Changes in definitions 66

10.2 What does the data look like? 66

10.2.1 Raw data 66

10.2.2 Cleaned data 66

10.3 What variables are in the data? 66

10.3.1 Key variables 66

10.3.2 Known issues with the data 66

10.4 Final thoughts 66

Chapter 1

Preface

If you've read an article about crime or arrests in the United States in the last half century, in most cases it was referring to the FBI's Uniform Crime Reporting Program Data, otherwise known as UCR data. UCR data is, with the exception of the more detailed data that only covers murders, a *monthly number of crimes or arrests reported to a single police agency* which is then gathered by the FBI into one file that includes all reporting agencies. Think of your home town. This data will tell you how many crimes were reported for a small number of crimes or how many people (broken down by age, sex, and race) were arrested for a (larger) set of crimes in that city (if the city has multiple police agencies, it will use the agency which is the primary agency on the case, usually the local police department) in a given month. This is a very broad measure of crime, and its uses in research - or understanding crime at all - is fairly limited. Yet is has become - and will likely remain among researchers for at least the next decade - the most important crime data in the United States.

UCR data is important for three reasons:

1. The definitions are standardized and all agencies follow them so you can compare across agencies and over time.
2. The data is available since 1960 so there is a long period of available data.
3. The data is available for most of the 18,000 police agencies in the United States so you can compare across agencies.

For most of this book we'll be discussing the caveats of the above reasons - or, more directly, why these assumptions are wrong - but these are the reasons why the data is so influential.

1.1 Motivation

By the end of each chapter you should have a firm grasp on the dataset the covered and how to use it properly. However, this book can't possibly cover every potential use case for the data so make sure to carefully examine the data for your own particular use. This benefits you because you'll know your data better and become a better research because of it. This benefits me because it'll increase the quality of research in my field.

I get a lot of emails from people asking questions about this data so part of my motivation for writing this book is to create a single place that answers as many questions as I can about the data. Again, the UCR data collection are the most commonly used crime datasets and there are still many current papers published with incorrect information about the data (including such simple aspects like what geographic unit data is in and what time unit it is in). So hopefully this book will decrease the number of misconceptions about this data, increasing overall research quality.[1]

1.2 Structure of the book

This remainder of this book will be divided into eight chapters: an intro chapter briefly summarizing each dataset and going over overall issues with UCR data, seven chapters each covering one of the seven UCR datasets, and a final one covering county-level data, a highly flawed but common use of the UCR data. I highly recommend that you start with the intro chapter (Chapter @ref(ucr_general)) since it covers topics that apply to each UCR dataset - such as understanding agency unique identifiers and why the data (correctly) has negative numbers - that won't be covered in individual chapters. Each of the UCR dataset chapters will follow the same format: we'll

[1] Ideally, this will also decrease the number of emails and increase the number of citations I receive.

start with a history of the data such as when it first because available and important changes in definitions or variables. We'll then look at which variables are available, and the most important (or most used) ones. The bulk of each chapter will be exploring the data to understand how reliable it is and what questions we can adequately answer using it.

Since manuals are boring, I'll try to include graphs and images to try to alleviate the boredom. That said, I don't think it's possible to make it too fun so sorry in advanced. This book is a mix of facts about the data, such as how many years are available, and my opinions about it, such as whether it is reliable. In cases of facts I'll just say a statement - e.g. "the offenses data is available since 1960". In cases of opinion I'll temper the statement by saying something like "in my opinion…" or "I think" or "I believe."

1.3 Citing this book

If this book was useful in your research, please cite it. To cite this book, please use the below citation:

Kaplan J (2021). *Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide*. https://github.com/jacobkap/ucrbook.

BibTeX format:

1.4 Sources of UCR data

There are a few different sources of UCR data available today. First, and probably most commonly used, is the data put together by the National Archive of Criminal Justice Data (NACJD)). This a team of out of the University of Michigan who manages a huge number of criminal justice datasets and makes them available to the public. If you have any questions about crime data - UCR or other data - I highly recommend you reach out to them for answers. They have a collection of data and excellent documentation available for UCR data available on their site here. One limitation to their data, however, is that each year of data is available as an individual file meaning that you'll need to concatenate each year together into a single file. Some years also have different column names (generally minor changes like

spelling robbery "rob" one year and "robb" the next) which requires more work to standardize before you could concatenate. They also only have data through 2016 which means that the most recent years (UCR data is available through 2019) of data are (as of this writing) unavailable.

Next, and most usable for the general public but limited to researchers, is the FBI's official website Crime Data Explorer. On this site you can chose an agency and see annual crime data (remember, UCR data is monthly so this isn't as detailed as it can be) for certain crimes. This is okay for the general public but only provides a fraction of the data available in the actual data so is really not good for researchers.

Finally, I have my own collection of UCR data available publicly on openICPSR, a site which allows people to submit their data for public access. For each of these datasets I've taken the raw data from the FBI (for early years of homicide data this is actually from NACJD since the FBI's raw data is wrong and can't be parsed. For later years of homicide data this is from the FBI's raw data.) and read it into R. Since the data is only available from the FBI as fixed-width ASCII files, I created a setup file (we'll explain exactly how reading in this kind of data works in the next chapter) and read the data and then very lightly cleaned the data (i.e. only removing extreme outliers like an agency having millions of arsons in a month). For each of these datasets I detail what I've done to the data and briefly summarize the data (i.e. a very short version of this book) on the data's page on openICPSR. The main advantage is that all my data has standard variable names and column names and, for data that is small enough, provide the data as a single file that has all years. For large datasets like the arrest data I break it down into parts of the data and not all years in a single file. The downside is that I don't provide documentation other than what's on the openICPSR page and only provide data in R and Stata format. I also have a similar site to the FBI's Crime Data Explorer but with more variables available, that site is available here.

It's worth mentioning a final source of UCR information. This is the annual Crimes in the United States report released by the FBI each year around the start of October.[2] As an example, here is the website for the 2019 report. In

[2]They also release a report about the first 6-months of the most recent year of data before the October release, but this is generally an estimate from a sample of agencies so is far less useful.

this report is summarized data which in most cases estimates missing data and provides information about national and subnational (though rarely city-level) crime data. As with the FBI's site it is only a fraction of the true data available, so is not a very useful source of crime data. Still, this is a very common source of information used by researchers.

1.5 Where to find the data used in this book

The data I am using in this book is the cleaned (we'll discuss in more detail exactly what I did to clean each dataset in the dataset's chapter, but the short answer is that I did very little.) and concatenated data that I put together from the raw data that the FBI releases. That data is available on my website here. I am hosting this book through GitHub which has a maximum file size allowed that is far smaller than these data so you'll need to go to my site to download the data, it's not available through this book's GitHub repo. For some examples I'm using the data before I cleaned it of outliers (as an example of the outliers present before I removed them) so that data is not publicly available.

1.6 NIBRS data

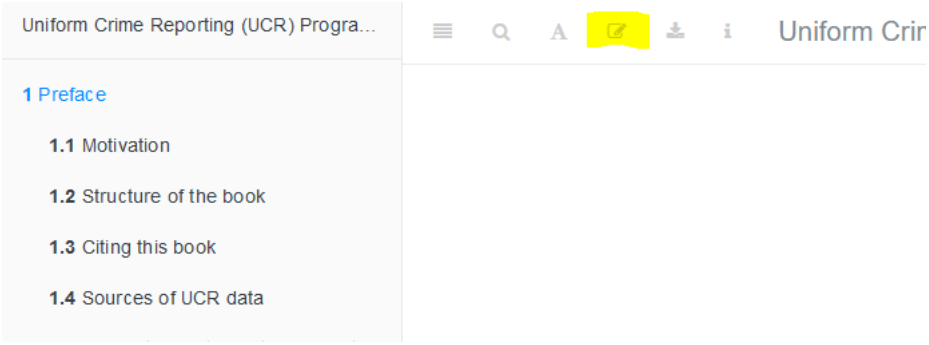
Another source of FBI data, and one sometimes considered part of the UCR data collection, is the National Incident-Based Reporting System (NIBRS) data. Like its name implies this is an incident-level dataset which has detailed information about each incident reported to the police, including incident circumstances, and victim and offender information. This is also the data that the FBI has declared will replace UCR data starting in 2021, meaning that they will no longer collect UCR data and only allow agencies to submit NIBRS data. NIBRS data is a complex and highly rich dataset that deserves its own book to really understand, so I will not be discussing it any further in this book.

1.7 This book ... in one paragraph

UCR data is complicated and riddled with limitations, so a naive use of it is certainly wrong. In my opinion, the vast majority of papers that use UCR data should have been rejected. They generally only briefly mention that the data used was UCR data and do not discuss issues with the data such as under-reporting, imputation procedures, or missing data. While their results may not change (or at least not change enough to affect the results) once they account for this, the lack of even bothering to discuss these issues is enough of an issue for an immediate desk reject. Surprisingly, this problem is even more severe for the worst UCR data, county-level imputed data and hate crime data, demonstrating that the authors who use that data are generally the ones least qualified (otherwise they wouldn't use it at all) to do so.

1.8 How to contribute to this book

If you have any questions, suggestions, or find any issues, please email me at jkkaplan6 [@] gmail.com. For more minor issues like typos or grammar mistakes, you can edit the book directly through its GitHub page. That'll make an update for me to accept, which will change the book to include your edit. To do that, click the edit button at the top of the site - the button is highlighted in the below figure. You will need to make a GitHub account to make edits. When you click on that button you'll be taken to a page that looks like a Word Doc where you can make edits. Make any edits you want and then scroll to the bottom of the page. There you can write a short (please, no more than a sentence or two) description of what you've done and then submit the changes for me to review.



About the author

Jacob Kaplan holds a PhD and a master's degree in criminology from the University of Pennsylvania and a bachelor's degree in criminal justice from California State University, Sacramento. His research focuses on Crime Prevention Through Environmental Design (CPTED), specifically on the effect of outdoor lighting on crime. He is the author of several R packages that make it easier to work with data, including fastDummies and asciiSetupReader. His website allows easy analysis of crime-related data and he has released over a dozen crime data sets (primarily FBI UCR data) on openICPSR that he has compiled, cleaned, and made available to the public.

For a list of papers he has written (including working papers), please see here.

For a list of data sets he has cleaned, aggregated, and made public, please see here.

For a list of R packages he has created, please see here.

Chapter 2

Overview of the Data

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: maps

## Skipping install of 'fiftystater' from a github remote, the SHA1 (28e7fa54)
##   Use `force = TRUE` to force installation

##
## -- Column specification --------------------------------------------------
## cols(
##   .default = col_double(),
##   agency = col_character(),
##   year = col_date(format = ""),
##   state = col_character(),
```

```
##   ORI = col_character()
## )
## i Use `spec()` for the full column specifications.
```

One of the first, and most important, questions I think people have about crime is a simple one: is crime going up? Answering it seems simple - you just count up all the crimes that happen in an area and see if that number of bigger than it was in previous times.

However, putting this into practice invites a number of questions, all of which affect how we measure crime. First, we need to define what a crime is? Not philosophically what actions are crimes - or what should be crimes - but literally which of the many thousands of different criminal acts (crimes as defined by state law) should be considered in this measure. Should murder count? Most people would say yes. How about jaywalking or speeding? Many would say probably not. Should marital rape be considered a crime? Now, certainly most people (all, I would hope) would say yes. But in much of the United States it wasn't a crime until the 1970s (Bennice and Resick, 2003; McMahon-Howard et al., 2009).

Next, we have to know what geographic and time unit to measure crimes at since these decisions determine how precise we can measure crime and when it changed. That is, if you are mugged on Jan 1st at exactly 12:15pm right outside your house, how do we record it? Should we be as precise as including the exact time and location (your home address or coordinates for the address)? Out of privacy concerns to the victim, should we only include a larger time unit (such as hour of the day or just the day without any time of day) or a larger geographic unit (such as a Census Tract or the city)?

The final question is that when a crime occurs, how do we know? That is, when we want to count how many crimes occurred do we ask people how often they've been victimized, do we ask people how often they commit a crime, do we look at crimes reported to police, crimes charged in a criminal court? Each of these measures will likely give different answers as to how many crimes occurred.[1]

[1]The Bureau of Justice Statistics does measure crime by asking a random sample of people whether they were the victim of a crime. For more on this, please see their National Crime Victimization Survey reports

The FBI answered all of these questions in 1929 when they began the Uniform Crime Reporting (UCR) Program Data, or UCR data for short. **Crime consists of eight crime categories - murder, rape, robbery, aggravated assault, burglary, motor vehicle theft, theft, and simple assault - that are reported to the police and is collected each month from each agency in the country.** These decisions, born primarily out of the resource limitations of 1929 (e.g. no computers), have had a major impact on criminology research. The first seven crime categories - known as "Index Crimes" or "Part 1 crimes" (or "Part I" sometimes) - are the ones used to measure crime in many criminology papers, even when the researchers have access to data that covers a broader selection of crimes than these.[2] The crime data actually also includes the final crime, simple assault, though it is not included as an index crime and is, therefore, generally ignored by researchers - a relatively large flaw in most studies that we'll discuss in more detail in Section @ref(index_crimes).

If you think that decisions made nearly a century ago are probably not the most useful for current research - and they're almost certainly not the decisions that you would have made - then you've struck at the core of this book. As researchers, we are relying on datasets whose creation was made so long ago that

2.1 Crimes included in the UCR datasets

UCR data covers only a subset - and for the crime data, a very small subset - of all crimes that can occur. It also only includes crimes reported to the police. So there are two levels of abstraction from a crime occurring to it being included in the data: a crime must occur that is one of the crimes included in the UCR collection (we detail all of these crimes below) *and* the victim must report the incident to the police. While the crimes included are only a small selection of crimes - which were originally chosen since at the time the UCR was designed these were considered important offenses and among the best reported - this is an important first step to understanding the data. In this section we'll go over the crimes included in the two main

[2] Arson is also an index crime but was added after these initial seven were chosen and is not included in the crimes dataset (though is available separately) so is generally not included in studies that use index crimes.

UCR datasets: Offenses Known and Clearances by Arrest and (which I like to call the "crime" dataset) and the Arrests by Age, Sex, and Race (the "arrests" dataset). These are the most commonly used UCR datasets and the stolen property and homicide datasets are simply more detailed subsets of these datasets. The hate crime dataset can cover a broader selection of offenses than in the crimes or arrests data, so we'll discuss those in the hate crimes chapter.

2.1.1 Crimes in the Offenses Known and Clearances by Arrest dataset

As mentioned above - and as most criminology papers will tell you - the crimes included in the UCR's Offenses Known and Clearances by Arrest data are the seven index crimes (eight when including arson, though arson is in its own dataset) - homicide, rape, robbery, aggravated assault, burglary, theft, and motor vehicle theft - as well as simple assault. This is true but incomplete. The data also includes subcategories for all crimes other than theft - though theft has its own UCR dataset which goes into detail about the thefts. Both robbery and aggravated assault, for example, have subcategories showing which weapon the offender used (if any) during the crime. This allows for a more detailed understanding of the crime than looking only at the broad category. I'm not sure why most research includes only the broader categories and doesn't tend to look at subcategories, but that seems to be the case in most studies that I have read. Some police agencies only report the broader categories and don't report subcategories, but most report subcategories so this is an under-exploited source of data.

1. Homicide

- Murder and non-negligent manslaughter
- Manslaughter by negligence

2. Rape

- Rape
- Attempted rape

3. Robbery

- With a firearm
- With a knife of cutting instrument
- With a dangerous weapon not otherwise specified
- Unarmed - using hands, fists, feet, etc.

4. Aggravated assault (assault with a weapon or causing serious bodily injury)

- With a firearm
- With a knife of cutting instrument
- With a dangerous weapon not otherwise specified
- Unarmed - using hands, fists, feet, etc.

5. Burglary

- With forcible entry
- Without forcible entry

- Attempted burglary with forcible entry

6. Theft (other than of a motor vehicle)
7. Motor vehicle theft

- Cars
- Trucks and buses
- Other vehicles

8. Simple Assault

2.1.2 Crimes in the Arrests by Age, Sex, and Race dataset

The crimes included in the Arrests by Age, Sex, and Race - the "arrest" data tells you how many people were arrested for a particular crime category - are different than those in the crime data. The arrest data covers a wider variety of crimes, including drug and alcohol crimes, gambling, and fraud. However, it is also less detailed than the crime dataset when it comes to violent crime. While it covers the same broad categories of violent crimes as the crimes data - murder, rape, robbery, aggravated assault, and simple assault - it doesn't include the more detailed breakdown that is available in the crime data. For example, in the crime data robbery is included as well as the subcategories of robbery with a gun, robbery with a knife, robbery with another dangerous weapon, and robbery without a weapon. In comparison, the arrest data only includes robbery without any subcategories.

1. Homicide

- Murder and non-negligent manslaughter

- Manslaughter by negligence

2. Rape
3. Robbery
4. Aggravated assault
5. Burglary
6. Theft (other than of a motor vehicle)
7. Motor vehicle theft
8. Simple assault
9. Arson
10. Forgery and counterfeiting
11. Fraud
12. Embezzlement
13. Stolen property - buying, receiving, and possessing
14. Vandalism
15. Weapons offenses - carrying, possessing, etc.
16. Prostitution and commercialized vice
17. Sex offenses - other than rape or prostitution
18. Drug abuse violations - total

- Drug sale or manufacturing
  - Opium and cocaine, and their derivatives (including morphine and heroin)
  - Marijuana
  - Synthetic narcotics
  - Other dangerous non-narcotic drugs
- Drug possession
  - Opium and cocaine, and their derivatives (including morphine and heroin)
  - Marijuana
  - Synthetic narcotics
  - Other dangerous non-narcotic drugs

19. Gambling - total

- Bookmaking - horse and sports

- Number and lottery
- All other gambling

20. Offenses against family and children - nonviolent acts against family members. Includes neglect or abuse, nonpayment of child support or alimony.
21. Driving under the influence (DUI)
22. Liquor law violations - Includes illegal production, possession (e.g. underage) or sale of alcohol, open container, or public use laws. Does not include DUIs and drunkenness.
23. Drunkenness - i.e. public intoxication
24. Disorderly conduct
25. Vagrancy - includes begging, loitering (for adults only), homelessness, and being a "suspicious person."
26. All other offenses (other than traffic) - a catch-all category for any arrest that is not otherwise specified in this list. Does not include traffic offenses. Very wide variety of crimes are included - use caution when using!
27. Suspicion - "Arrested for no specific offense and released without formal charges being placed."
28. Curfew and loitering law violations - for minors only.
29. Runaways - for minors only.

2.1.3 What is an index (or part 1) crime?

One of the first (and seemingly last) thing that people tend to learn about UCR crime data is that it covers something called an "index crime."[3] Index crimes, sometimes written as Part 1 or Part I crimes, are the seven crimes originally chosen by the FBI to be included in their measure of crimes as these offenses were both considered serious and generally well-reported so would be a useful measure of crime. Index crimes are often broken down into property index crimes - burglary, theft, and motor vehicle theft (and arson now, though that's often not included and is poorly reported) - and violent index crimes (murder, rape, robbery, and aggravated assault). The "index"

[3]Index crimes are sometimes capitalized as "Index Crimes" though I've seen it written both ways. In this book I keep it lowercase as "index crimes."

is simply that all of the crimes are summed up into a total count of crimes (violent, property, or total) for that police agency.

The biggest problem with index crimes is that it is simply the sum of 8 (or 7 since arson data usually isn't available) crimes. Index crimes have a huge range in their seriousness - it includes, for example, both murder and theft - so summing up the crimes gives each crime equal weight. This is clearly wrong as 100 murders is more serious than 100 thefts. This is especially a problem as less serious crimes (theft mostly) are far more common than more serious crimes. In 2017, for example, there were 1.25 million violent index crimes in the United States. That same year had 5.5 million thefts. So using index crimes as your measure of crimes undercounts the seriousness of crimes. Looking at total index crimes is, in effect, mostly just looking at theft. Looking at violent index crimes mostly measures aggravated assault. This is especially a problem because it hides trends in violent crimes. San Francisco, as shown in Figure **?**(fig:sfThefts), has had a huge increase in index crimes in the last several years. When looking closer, that increase is driven almost entirely by the near doubling of theft since 2011. During the same years, index violent crimes have stayed fairly steady. So the city isn't getting more dangerous - at least not in terms of violent index crimes increasing - but it appears like it is due to just looking at total index crimes.

While many researchers divide index crimes into violent and nonviolent categories, which helps but even this isn't entirely sufficient. Take Chicago as an example. It is a city infamous for its large number of murders. But as a fraction of index crimes, Chicago has a rounding error worth of murders. Their 653 murders in 2017 is only 0.5% of total index crimes. For violent index crimes, murder makes up 2.2%. As seen in Figure 2.2, in no year where data is available did murders account for more than 3.5% of violent index crimes; and, while murders are increasing as a percent of violent index crimes they still account for no more than 2.5% in most years. What this means is that changes in murder are very difficult to detect. If Chicago had no murders this year, but a less serious crime (such as theft) increased slightly, we couldn't tell from looking at the number of index crimes, even from violent index crimes. As discussed in the below section, using this sample of crimes as the primary measure of crimes - and particularly of violent crimes - is also misleading as it excludes important - and highly common relative to index crimes - offenses, such as simple assault.

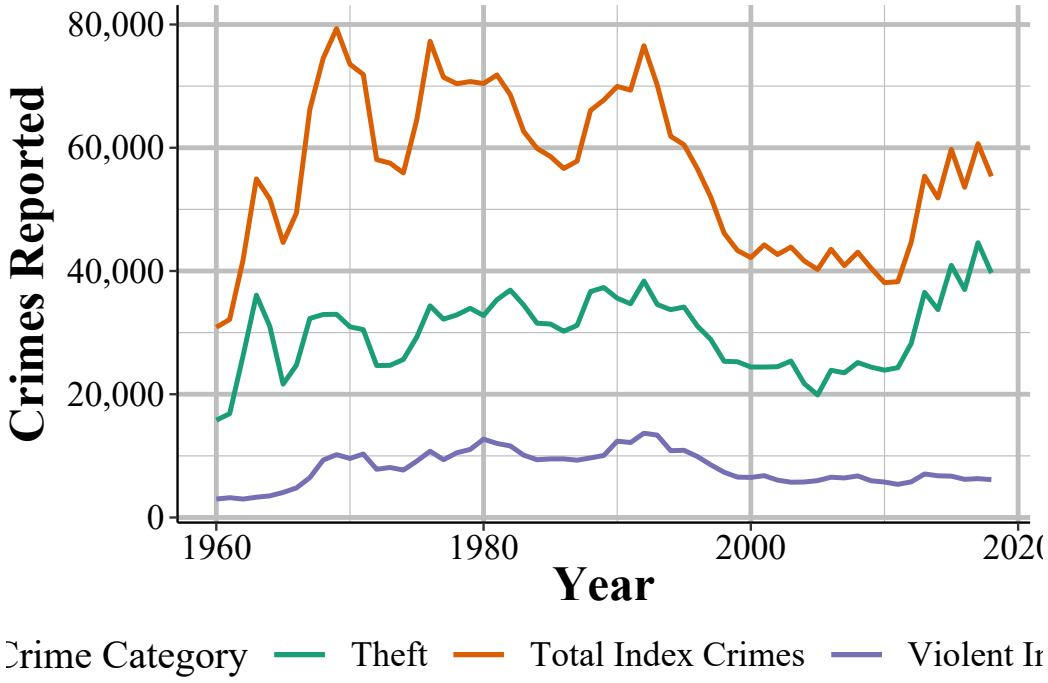


Figure 2.1: Thefts and total index crimes in San Francisco, 1960-2018.

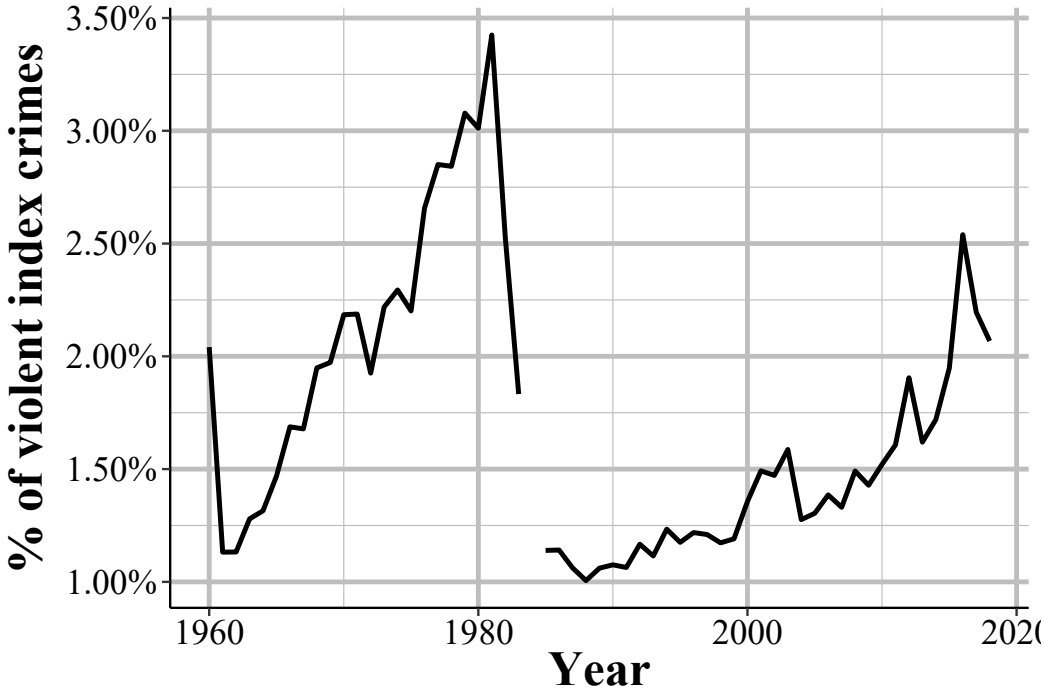


Figure 2.2: Murders in Chicago as a percent of violent index crimes, 1960-2018.

2.1.3.1 What is a violent crime?

An important consideration in using this data is defining what a "violent crime" is. One definition, and the one that I think makes the most sense, is that a violent crime is one that uses force or the threat of force. For example, if I punch you in the face, that is a violent crime. If I stab you, that is a violent crime. While clearly different in terms of severity, both incidents used force so I believe would be classified as a violent crime. The FBI, and most researchers, reporters, and advocates would disagree. Organizations ranging from the FBI itself to Pew Research Center and advocacy groups like the Vera Institute of Justice and the ACLU all define the first examine as a non-violent crime and the second as a violent crime. They do this for three main reasons.

The first reason is that simple assault is not an index crime, so they don't include it when measuring violent crime. It is almost a tautology in criminology that you use index crimes as the measure of crime since that's just what people do. As far as I'm aware, this is really the main reason why researchers justify using index crimes: because people in the past used it so that's just what is done now.[4] This strikes me as a particularly awful way of doing anything, more so since simple assault data has been available almost as long as index crime data.[5]

The second reason - and one that I think makes sense for reporters and advocates who are less familiar with the data, but should be unacceptable to researchers - is that people don't know that simple assault is included, or at least don't have easy access to it. Neither the FBI's annual report page not their official crime data tool website include simple assault since they only include index crimes. For people who rely only on these sources - and given that using the data itself requires at least some programming skills, I think this accounts for most users and certainly nearly all non-researchers - it is not possible to access simple assault crime data (arrest data does include simple assault on these sites).

The final reason is that it benefits some people's goals to classify violent crime as only including index crimes. This is because simple assault is extremely

[4] I've also seen the justification that aggravated assault is more serious since it uses a weapon, though as the UCR subcategory of aggravated assault without a weapon clearly shows, aggravated assault does not require use of a weapon.

[5] Simple assault is first available in 1964. Index crime data is available since 1960.

common compared to violent index crimes - in many cities simple assault is more common than all violent index crimes put together - so excluding simple assault makes it seem like fewer arrests are violent than they are when including simple assault. For example, a number of articles have noted that marijuana arrests are more common than violent crime arrests (Ingraham, 2016; Kertscher, 2019; Devito, 2020; Earlenbaugh, 2020; Edwards et al., 2020) or that violent crime arrests are only 5% of all arrests (Neusteter and O'Toole, 2019; Speri, 2019).[6] While true when considering only violent index crimes, including simple assault as a violent crime makes these statements false. Some organizations call the violent index crimes "serious violent crimes" which is an improvement but even this is a misnomer since simple assault can lead to more serious harm than aggravated assault. An assault becomes aggravated if using a weapon or there is the *potential* for serious harm, even if no harm actually occurs.[7]

As an example of this last point, 2.3 shows the number of violent index crimes and simple assaults each year from 1960 to 2018 in Houston, Texas (simple assault is not reported in UCR until 1964, which is why 1960-1963 show zero simple assaults). In every year where simple assault is reported, there are more simple assaults than aggravated assaults. Beginning in the late 1980s, there are also more simple assaults than total violent index crimes. Excluding simple assault from the being a violent crime greatly underestimates violence in the country.

2.2 Issues common across UCR datasets

In this section we'll discuss issues common to most or all of the UCR datasets. For some of these, we'll come back to the issues in more detail in the chapter for the datasets most affected by the problem.

[6] The FBI's report for arrests does include simple assault so the second reason people may not include it does not apply here.

[7] UCR data provides no information about the harm caused to victims. The new FBI dataset NIBRS actually does have a variable that includes harm to the victim so if you're interested in measuring harm (an understudied topic in criminology), that is the dataset to use.

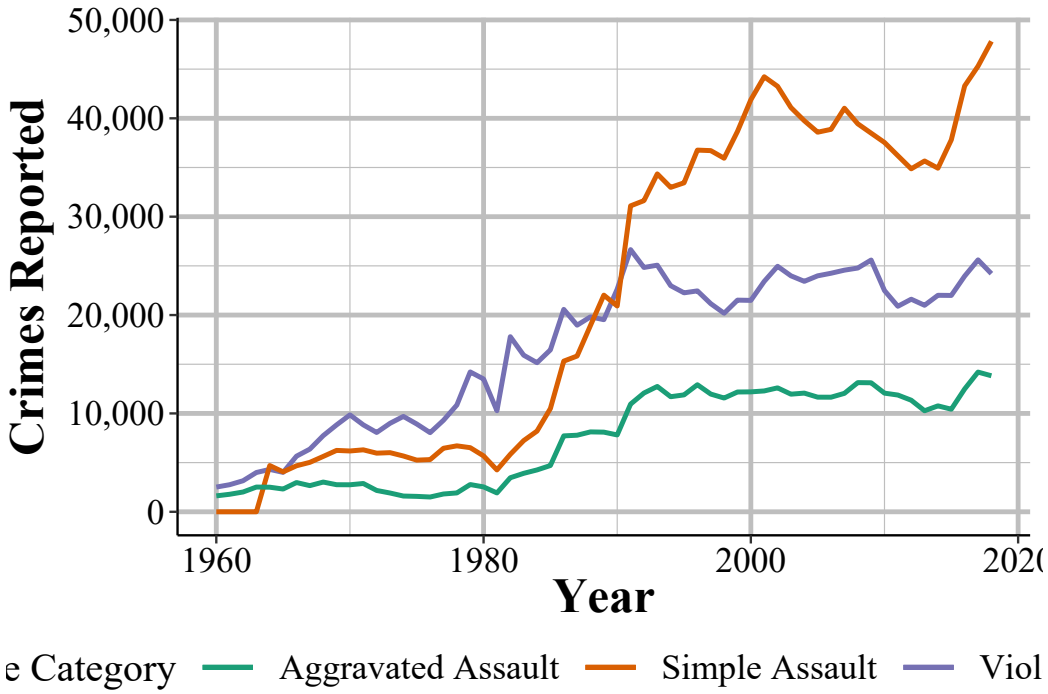


Figure 2.3: Reported crimes in Philadelphia from 1960 to 2018. Violent index crimes are aggravated assault, rape, robbery, and murder.

2.2.1 Negative numbers

One of the most common questions people have about this data is why there are negative numbers, and if they are a mistake. Negative numbers are not a mistake. The UCR data is monthly so police agencies report the number of monthly crimes that are known to them, either reported to them or discovered by the police. In some cases the police will determine that a crime reported to them didn't actually occur - which they called an "unfounded crime." An example that the FBI gives for this is a person reports their wallet stolen (a theft) and then later finds it, so a crime was initially reported but no crime actually occurred.

How this works when the police input the data is that an unfounded crime is reported both as an unfounded crime and as a negative actual crime - the negative is used to zero out the erroneous report of the actual crime. So the report would look like 1 actual theft (the crime being reported), -1 theft (the crime being discovered as not have happened), and 1 unfounded theft. If both incidents occurred in the same month then this would simply be a single unfounded theft occurring, with no actual theft - literally a value of 1 + -1 = 0 thefts.

Negative values occur when the unfounding happens in a later month than the crime report. In the theft case, let's say the theft occurred in January and the discovery of the wallet happens in August. Assuming no other crimes occurred, January would have 1 theft, and August would have -1 thefts and 1 unfounded theft. There is no way of determining in which month (or even which year) an unfounded crime was initially reported in. When averaging over the long term, there shouldn't be any negative numbers as the actual and unfounded reports will cancel themselves out. However, when looking at monthly crimes - particularly for rare crimes - you'll still see negative numbers for this reason. Since crimes can be unfounded for reports in previous years, you can actually see entire year's crime counts be negative, though this is much rarer than monthly values.[8]

[8]From 1960-2018, there were 39 agency-years with a negative count of murders.

2.2.2 Agency population value

Each of the UCR datasets include a population variable that has the estimated population under the jurisdiction of that agency.[9] This variable is often used to create crime rates that control for population. In cases where jurisdiction overlaps, such as when a city has university police agencies or county sheriff's in counties where the cities in that county have their own police, UCR data assigns the population covered to the most local agency and zero population to the overlapping agency. So an agency's population is the number of people in that jurisdiction that isn't already covered by a different agency.

For example, in the city of Los Angeles in California has nearly four million residents according to the US Census. There are multiple police agencies in the city, including the Los Angeles Police Department, the Los Angeles County Sheriff's Office, the California Highway Patrol that operates in the area, airport and port police, and university police departments. If each agency reported the number of people in their jurisdiction - which all overlap with each other - we would end up with a population far higher than LA's four million people. To prevent double-counting population when agency's jurisdictions overlap, the non-primary agency will report a population of 0, even though they still report crime data like normal. As an example, in 2018 California State University - Los Angeles reported 92 thefts and a population of 0. Those 92 thefts are not counted in the Los Angeles Police Department data, even though the department counts the population. To get complete crime counts in Los Angeles, you'd need to add up all police agencies within in the city; since the population value is 0 for non-LAPD agencies, both the population and the crime sum will be correct.

The UCR uses this method even when only parts of a jurisdiction overlaps. Los Angeles County Sheriff has a population of about one million people, far less than the actual county population (the number of residents, according to the Census) of about 10 million people. This is because the other nine million people are accounted for by other agencies, mainly the local police agencies in the cities that make up Los Angeles County.

[9]Jurisdiction here refers to the boundaries of the local government, not any legal authority for where the officer can make arrests. For example, the Los Angeles Police Department's jurisdiction in this case refers to crimes that happen inside the city or are otherwise investigated by the LAPD - and are not primarily investigated by another agency.

The population value is the population who reside in that jurisdiction, and does not count people who are in the area but don't live there, such as tourists or people who commute there for work. This means that using the population value to determine a rate can be misleading as some places have much higher numbers of non-residents in the area (e.g. Las Vegas, Washington D.C.) than others.

2.2.3 Reporting is voluntary ... so some agencies don't (or report partially) {voluntary}

When an agency reports their data to the FBI, they do so voluntarily - there is no nationally requirement to report.[10] This means that there is inconsistency in which agencies report, how many months of the year they report for, and which variables they include in their data submissions. This can lead

In general, more agencies report their data every year and once an agency begins reporting data they tend to keep reporting. The UCR datasets are a collection of separate, though related, datasets and an agency can report to as many of these datasets as they want - an agency that reports to one dataset does not mean that they report to other datasets. Figure **??** shows the number of agencies that submitted at least one month of data - based on a highly flawed measure that we'll discuss in Section **??** - to the Offenses Known and Clearances by Arrest data in the given year. For the first decade of available data under 8,000 agencies reported data and this grew to over 13,500 by the late 1970s before plateauing for about a decade. The number of agencies that reported their data actually declined in the 1990s, driven primarily by many Florida agencies temporarily dropping out, before growing steadily to nearly 17,000 agencies in 2010; from here it kept increasing but slower than before.

There are approximately 18,000 police agencies in the United States so recent data has reports from nearly all agencies, while older data has far fewer agencies reporting. For earlier data, however, you're dealing with a smaller share of agencies meaning that you have a large amount of missing data and a less representative sample.

[10] Some states do mandate that their agencies report, but this is not always followed.

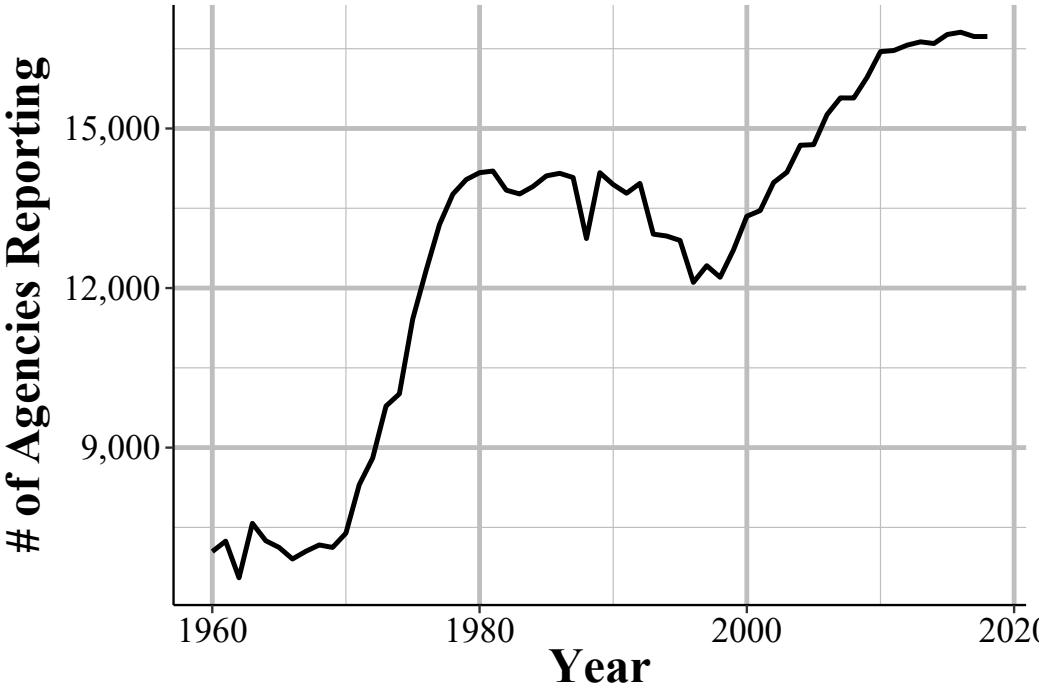


Figure **??** repeats this figure but now including only agencies with 100,000 people or more in their jurisdiction. While these agencies have a far more linear trend than all agencies, the basic lesson is the same: recent data has most agencies reporting; old data excludes many agencies.

In 2003 the New York Police Department stopped reporting this category of offense, resuming only in 2013. They continued to report the broader aggravated assault category, but not any of the subsections of aggravated assaults which differentiate the weapon used.

2.2.3.1 Number of months reported {monthsReported}

2.2.4 Zero crimes vs no reports

When an agency does not report, we see it in the data as reporting zero crimes, not reporting NA or any indicator that they did not report.

As discussed in Section **??**, agencies voluntarily report, and in some cases doesn't report or don't report certain months or crimes.

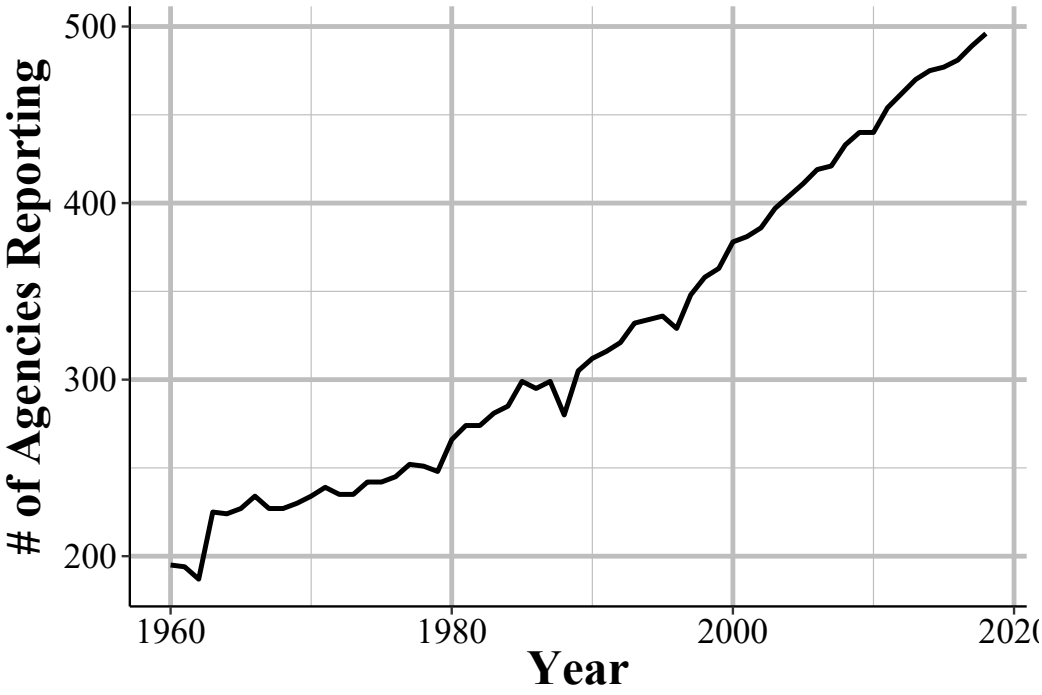


Figure 2.4: The annual number of agencies with a population of 100,000 or higher reporting to the Offenses Known and Clearances by Arrest dataset. Reporting is based on the agency reporting at least one month of data in that year.

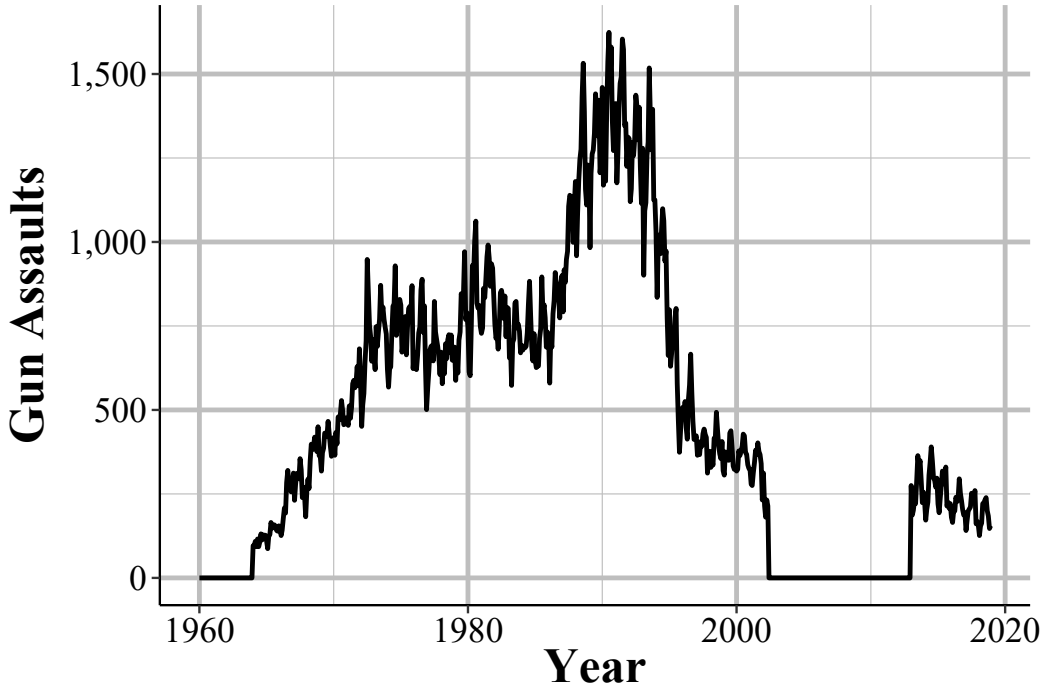


Figure 2.5: Monthly reports of gun assaults in New York City, 1960-2018.

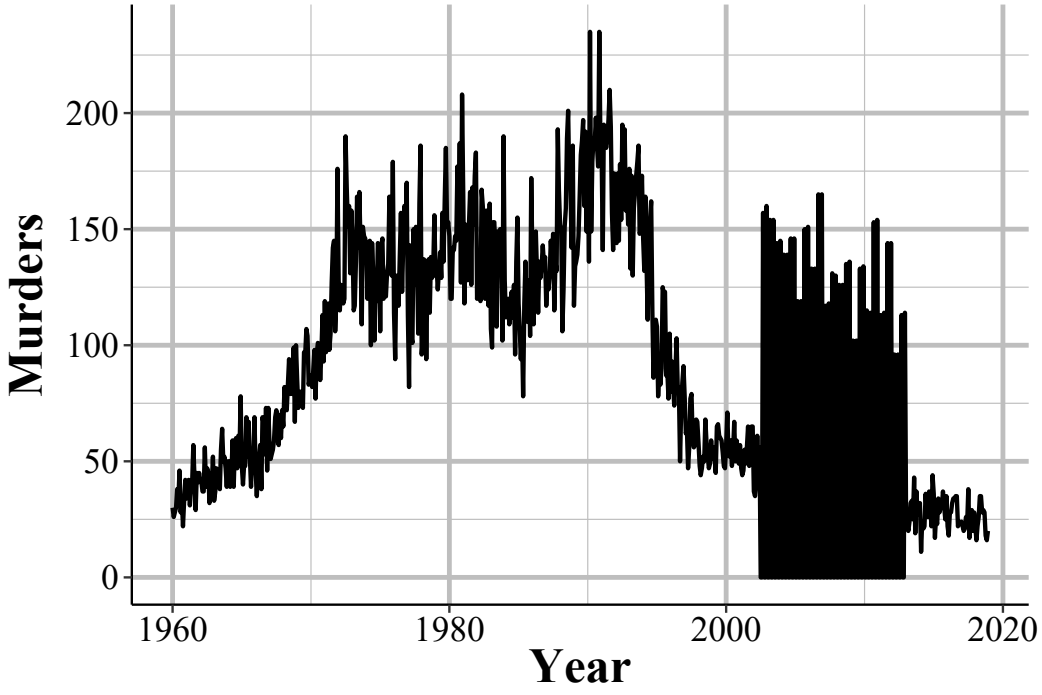


Figure 2.6: Monthly murders in New York City, 1960-2018. During the 2000s, the police department began reporting quarterly instead of monthly and then resumed monthly reporting.

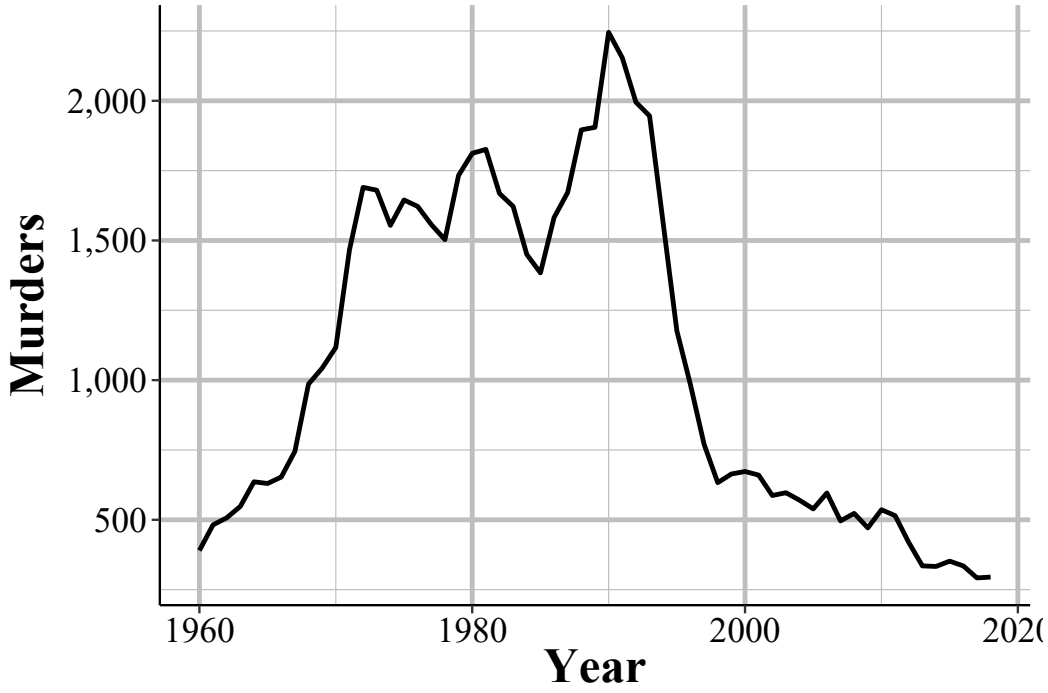


Figure 2.7: Annual murders in New York City, 1960-2018.

In some cases it is clear when an agency stops reporting

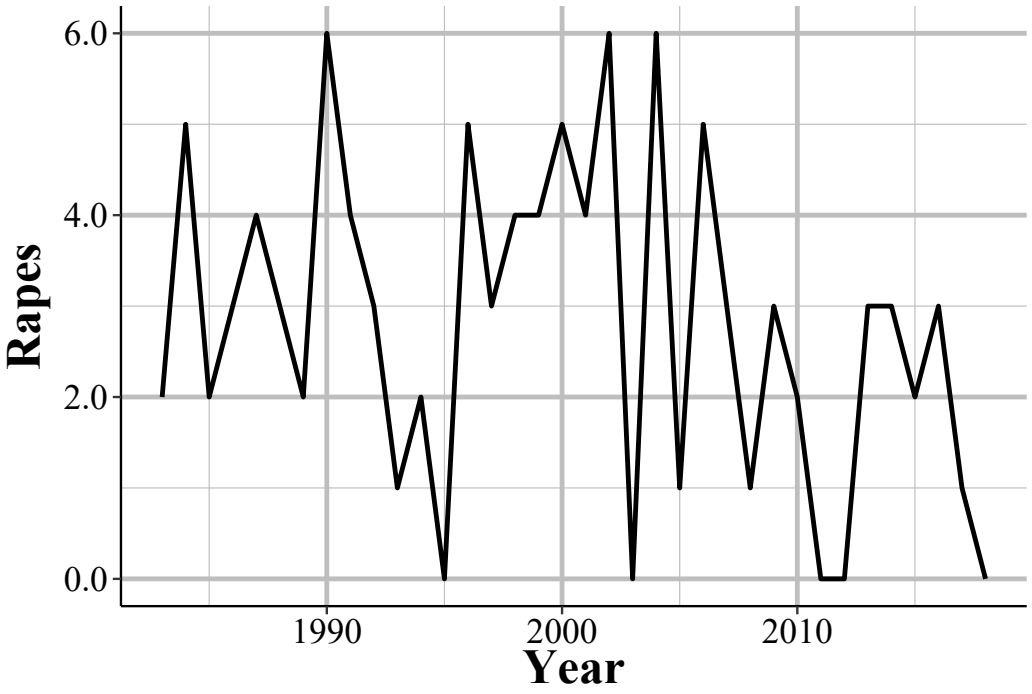


Figure 2.8: Annual rapes reported in Danville, CA, 1960-2018.

2.3 A summary of each UCR dataset

UCR data can be roughly summarized into two groups: crime data and arrest data. While there are several datasets included in the UCR data collection, they are all extensions of one of the above groups. For arrest data, you have information about who (by race and by age-gender, but not by race-gender or race-age other than within race you know if the arrestee is an adult or a juvenile). For crime data, you have monthly counts of a small number of crimes (many fewer than crimes covered in the arrest data) and then more specialized data on a subset of these crimes - information on homicides, hate crimes, assaults or deaths of police officers, and stolen property.

Each of these datasets will have its own chapter in this book where we discuss

the data thoroughly. Here is a very brief summary of each dataset which will help you know which one to use for your research. I still recommend reading that data's chapter since it covers important caveats and uses (or misuses) of the data that won't be covered below.

2.3.1 Offenses Known and Clearances by Arrest

The Offenses Known and Clearances by Arrest dataset - often called Return A, "Offenses Known" or, less commonly, OKCA - is the oldest and most commonly used dataset and measures crimes reported to the police. For this reason it is used as the main measure of crime in the United States and I tend to call it the "crimes dataset." This data includes the monthly number of crimes reported to the police or otherwise known to the police (e.g. discovered while on patrol) for a small selection of crimes, as well as the number of crimes cleared by arrest or by "exceptional means" (a relatively flawed and manipulable measure of whether the case is "solved"). It also covers the number reported but found by police to have not occurred. Since this data has monthly agency-level crime information it is often used to measure crime trends between police agencies and over time. The data uses something called a Hierarchy Rule which means that in incidents with multiple crimes, only the most serious is recorded - though in practice this affects only a small percent of cases, and primarily affects property crimes.

2.3.2 Arrests by Age, Sex, and Race

The Arrests by Age, Sex, and Race dataset - often called ASR or the "arrests data" - includes the monthly number of arrests for a variety of crimes and, unlike the crime data, breaks down this data by age and gender. This data includes a broader number of crime categories than the crime dataset (the Offenses Known and Clearances by Arrest data) though is less detailed on violent crimes since it does not breakdown aggravated assault or robberies by weapon type as the Offenses Known data does. For each crime it says the number of arrests for each gender-age group with younger ages (15-24) showing the arrestee's age to the year (e.g. age 16) and other ages grouping years together (e.g. age 25-29, 30-34, "under 10"). It also breaks down arrests by race-age by including the number of arrestees of each race (American Indian,

Asian, Black, and White) are the only included races) and if the arrestee is a juvenile (<18 years old) or an adult. The data does technically include a breakdown by ethnicity-age (e.g. juvenile-Hispanic, juvenile-non-Hispanic) but almost no agencies report this data (for many years zero agencies report ethnicity) so in practice the data does not include ethnicity. As the data includes counts of arrestees, people who are arrested multiple times are included in the data multiple times - it is not a measure of unique arrestees.

2.3.3 Law Enforcement Officers Killed and Assaulted (LEOKA)

The Law Enforcement Officers Killed and Assaulted data, often called just by its acronym LEOKA, has two main purposes.[11] First, it provides counts of employees employed by each agency - broken down by if they are civilian employees or sworn officers, and also broken down by gender. And second, it measures how many officers were assaulted or killed (including officers who die accidentally such as in a car crash) in a given month. The assault data is also broken down into shift type and type (e.g. alone, with a partner, on foot, in a car, etc.), the offender's weapon, and type of call they are responding to (e.g. robbery, disturbance, traffic stop). The killed data simply says how many officers are killed feloniously (i.e. murdered) or died accidentally (e.g. car crash) in a given month. The employee information is at the year-level so you know, for example, how many male police officers were employed in a given year at an agency, but don't know any more than that such as how many officers were on patrol, were detectives, were in special units, etc. This dataset is commonly used as a measure of police employees and is a generally reliable - though imperfect as we'll see - measure of how many police are employed by a police agency. The second part of this data, measuring assaults and deaths, is more flawed with missing data issues and data error issues (e.g. more officers killed than employed in an agency).

2.3.4 Supplementary Homicide Reports (SHR)

The Supplementary Homicide Reports dataset - often abbreviated to SHR - is the most detailed of the UCR datasets and provides information about

[11] This data is also sometimes called the "Police Employees" dataset.

the circumstances and participants (victim and offender demographics and relationship status) for homicides.[12] For each homicide incident it tells you the age, gender, race, and ethnicity of each victim and offender as well as the relationship between the first victim and each of the offenders (but not the other victims in cases where there are multiple victims). It also tells you the weapon used by each offender and the circumstance of the killing, such as a "lovers triangle" or a gang-related murder. As with other UCR data, it also tells you the agency it occurred in and the month and year when the crime happened.

2.3.5 Hate Crime Data

This dataset covers crimes that are reported to the police and judged by the police to be motivated by hate More specifically, they are, first, crimes which were, second, motivated - at least in part - by bias towards a certain person or group of people because of characteristics about them such as race, sexual orientation, or religion. The first part is key, they must be crimes - and really must be the selection of crimes that the FBI collects for this dataset. Biased actions that don't meet the standard of a crime, or are of a crime not included in this data, are not considered hate crimes. For example, if someone yells at a Black person and uses racial slurs against them, it is clearly a racist action. For it to be included in this data, however, it would have to extend to a threat since "intimidation" is a crime included in this data but lesser actions such as simply insulting someone is not included. For the second part, the bias motivation, it must be against a group that the FBI includes in this data. When this data collection began crimes against transgender people were not counted so if a transgender person was assaulted or killed because they were transgender, this is not a hate crime recorded in the data.[13]

So this data is really a narrower measure of hate crimes than it might seem. In practice it is (some) crimes motivated by (some) kinds of hate that are reported to the police. It is also the most under-reported UCR dataset with most agencies not reporting any hate crimes to the FBI. This leads

[12] If you're familiar with the National Incident-Based Reporting System (NIBRS) data that is replacing UCR, this dataset is the closest UCR data to it, though it is still less detailed than NIBRS data.

[13] The first year where transgender as a group was a considered a bias motivation was in 2014.

to huge gaps in the data with some states having extremely few agencies reporting crime - see, for example Figure 2.9 for state-level hate crimes in 2018 - agencies reporting some bias motivations but not others, agencies reporting some years but not others. While these problems exist for all of the UCR datasets, it is most severe in this data. This problem is exacerbated by hate crimes being rare even in agencies that report them - with such rare events, even minor changes in which agencies report or which types of offenses they include can have large effects.

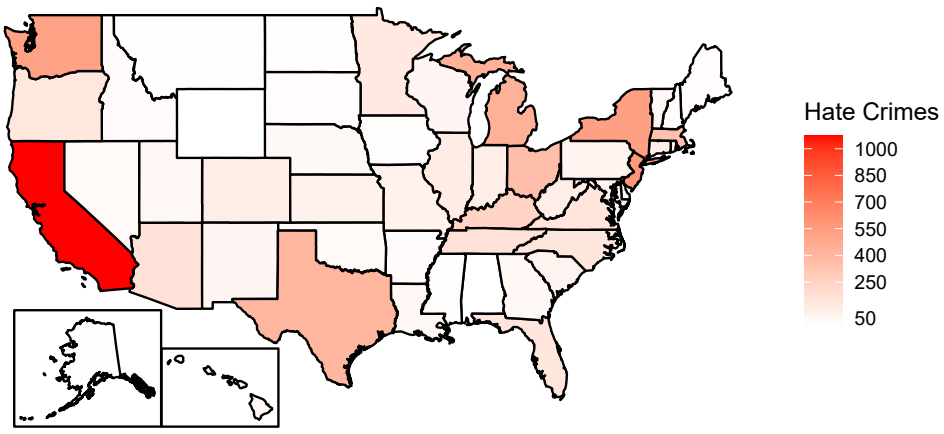


Figure 2.9: Total hate crimes by state, 2018.

2.3.6 Property Stolen and Recovered (Supplement to Return A)

The Property Stolen and Recovered data - sometimes called the Supplement to Return A (Return A being another name for the Offenses Known and Clearances by Arrest dataset, the "crime" dataset) - provides monthly

information about property-related offenses (theft, motor vehicle theft, robbery, and burglary), including the location of the offense (in broad categories like "gas station" or "residence), what was stolen (e.g. clothing, livestock, firearms), and how much the stolen items were worth.[14] The"recovered" part of this dataset covers the type and value of property recovered so you can use this, along with the type and value of property stolen, to determine what percent and type of items the police managed to recover. Like other UCR datasets this is at the agency-month level so you can, for example, learn how often burglaries occur at the victim's home during the day, and if that rate changes over the year or differs across agencies. The data, however, provides no information about the offender or the victim (other than if the victim was an individual or a commercial business [based on the location of the incident - "bank", "gas station", etc]). The value of the property stolen is primarily based on the victim's estimate of how much the item is worth (items that are decreased in value once used - such as cars - are supposed to be valued at the current market rate, but the data provides no indication of when it uses the current market rate or the victim's estimate) so it should be used as a very rough estimate of value.

2.3.7 Arson

The arson dataset provides monthly counts at the police agency-level for arsons that occur, and includes a breakdown of arsons by the type of arson (e.g. arson of a person's home, arson of a vehicle, arson of a community/public building) and the estimated value of the damage caused by the arson. This data includes all arsons reported to the police or otherwise known to the police (e.g. discovered while on patrol) and also has a count of arsons that lead to an arrest (of at least one person who committed the arson) and reports that turned out to not be arsons (such as if an investigation found that the fire was started accidentally). For each type of arson it includes the number of arsons where the structure was uninhabited or otherwise not in use, so you can estimate the percent of arsons of buildings which had the potential to harm people. This measure is for structures where people normally did not inhabit the structure - such as a vacant building where no one lives. A home where no one is home *at the time of the arson* does not count as an

[14] It also includes the value of items stolen during rapes and murders, if anything was stolen.

uninhabited building. In cases where the arson led to a death, that death would be recorded as a murder on the Offenses Known and Clearances by Arrest dataset - but not indicated anywhere on this dataset. If an individual who responds to the arson dies because of it, such as a police officer or a firefighter, this is not considered a homicide (though the officer death is still included in the Law Enforcement Officers Killed and Assaulted data) unless the arsonists intended to cause their deaths. Even though the UCR uses the Hierarchy Rule, where only the most serious offense that occurs is recorded, all arsons are reported - arson is except fro the Hierarchy Rule.

2.4 How to identify a particular agency (ORI codes)

In the UCR and other FBI data sets, agencies are identified using **OR**iginating Agency **I**dentifiers or an ORI. An ORI is a unique ID code used to identify an agency.[15] If we used the agency's name we'd end up with some duplicates since there can be multiple agencies in the country (and in a state, those this is very rare) with the same name. For example, if you looked for the Philadelphia Police Department using the agency name, you'd find both the "Philadelphia Police Department" in Pennsylvania and the one in Mississippi. Each ORI is a 7-digit value starting with the state abbreviation (for some reason the FBI incorrectly puts the abbreviation for Nebraska as NB instead of NE) followed by 5 numbers.[16] When dealing with specific agencies, make sure to use the ORI rather than the agency name to avoid any mistakes. In this book when discussing a particular agency I'll generally refer to it both by name and by ORI, but in the code used to generate figures and tables I identify that agency through their ORI code.[17]

For an easy way to find the ORI number of an agency, use this page on my site. Type an agency name or an ORI code into the search section and it will return everything that is a match.

[15] I will refer to this an an "ORI", "ORI code", and "ORI number", all of which mean the same thing.

[16] In the NIBRS data (another FBI data set) the ORI uses a 9-digit code - expanding the 5 numbers to 7 numbers.

[17] Since this isn't a programming book I won't include the code.

2.5 The data as you get it from the FBI

We'll finish this overview of the UCR data by briefly talking about format of the data that is released by the FBI, before the processing done by myself or NACJD that converts the data to a type that software like R or Stata or Excel can understand. The FBI releases their data as fixed-width ASCII files which are basically just an Excel file but with all of the columns squished together. As an example, below is the data as you receive it from the FBI for the Offenses Known and Clearances by Arrest dataset for 1960, the first year with data available. In the figure, it seems like there are multiple rows but that's just because the software that I opened the file in isn't wide enough - in reality what is shown is a single row that is extremely wide because there are over 1,500 columns in this data. If you scroll down enough you'll see the next row, but that isn't shown in the current image. What is shown is a single row with a ton of columns all pushed up next to each other. Since all of the columns are squished together (the gaps are just blank spaces because the value there is a space, but that doesn't mean there is a in the data. Spaces are possible values in the data and are meaningful), you need some way to figure out which parts of the data belong in which column.

The "fixed-width" part of the file type is how this works (the ASCII part basically means it's a text file). Each row is the same width - literally the same number of characters, including blank spaces. So you must tell the software you are using to process this file - by literally write code in something called a "setup file" but is basically just instructions for whatever software you use (R, SPSS, Stata, SAS can all do this) - which characters are certain columns. For example, in this data the first character says which type of UCR data it is (1 means the Offenses Known and Clearances by Arrest data) and the next two characters (in the setup file written as 2-3 since it is characters 2 through 3 [inclusive]) are the state number (01 is the state code for Alabama). So we can read this row as the first column indicating it is an Offenses Known data, the second column indicating that it is for the state of Alabama, and so on for each of the remaining columns. To read in this data you'll need a setup file that covers every column in the data (some software, like R, can handle just reading in the specific columns you want and don't need to include every column in the setup file).

The second important thing to know about reading in a fixed-width ASCII

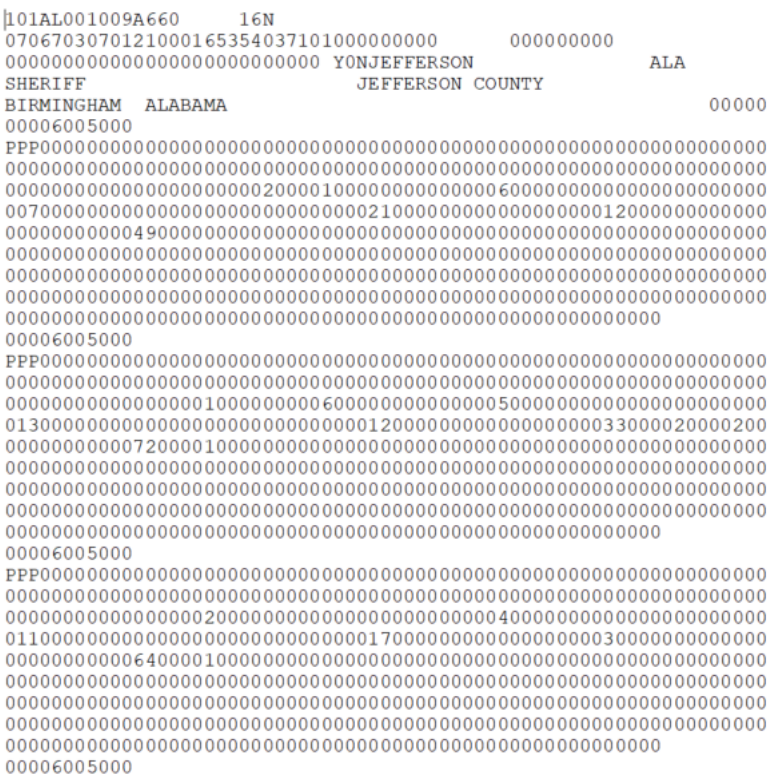


Figure 2.10: Fixed-width ASCII file for the 1960 Offenses Known and Clearances by Arrest dataset

file is something called a "value label."[18] For example, in the above image we saw the characters 2-3 is the state and in the row we have the value "01" which means that the state is "Alabama." Since this type of data is trying to be as small as efficient as possible, it often replaces longer values with shorter one and provides a translation for the software to use to convert it to the proper value when reading it. "Alabama" is more characters than "01" so it saves space to say "01" and just replace that with "Alabama" later on. So "01" would be the 'value" and "Alabama" would be the 'label' that it changes to once read.

Fixed-width ASCII files may seem awful to you reading it today, and it is awful to use. But it appears to be an efficient way to store data back many decades ago when data releases began but now is extremely inefficient - in terms of speed, file size, ease of use - compared to modern software so I'm not sure why they *still* release data in this format. But they do, and even the more *modern* (if starting in 1991, before I was born, is modern!) NIBRS data comes in this format. For you, however, the important part to understand is not how exactly to read this type of data, but to understand that people who made this data publicly available (such as myself and the team at NACJD) must made this conversion process.[19] **This conversion process, from fixed-width ASCII to a useful format is the most dangerous step taken in using this data - and one that is nearly entirely unseen by researchers.**

Every line of code you write (or, for SPSS users, click you make) invites the possibility of making a mistake.[20] The FBI does not provide a setup file with the fixed-width ASCII data so to read in this data you need to make it yourself. Since some UCR data are massive, this involves assigning the column width for thousands of columns and the value labels for hundreds of

[18]For most fixed-width ASCII files there are also missing values where it'll have placeholder value such as -8 and the setup file will instruct the software to convert that to NA. UCR data, however, does not have this and does not indicate when values are missing in this manner.

[19]For those interested in reading in this type of data, please see my R package asciiSetupReader.

[20]Even highly experienced programmers who are doing something like can make mistakes. For example, if you type out "2+2" 100 times - something extremely simple that anyone can do - how often will you mistype a character and get a wrong result? I'd guess that at least once you'd make a mistake.

different value labels.[21] A typo anywhere could have potentially far-reaching consequences, so this is a crucial weakpoint in the data cleaning process - and one in which I have not seen anything written about before. While I have been diligent in checking the setup files and my code to seek out any issues - and I know that NACJD has a robust checking process for their own work - that doesn't mean our work is perfect.[22] Even with perfection in processing the raw data to useful files, decisions we make (e.g. what level to aggregate to, what is an outlier) can affect both what type of questions you can ask when using this data, and how well you can answer them.

[21] With the exception of the arrest data and some value label changes in hate crimes and homicide data, the setup files remain consistent you a single file will work for all years for a given dataset. You do not need to make a setup file for each year.

[22] For evidence of this, please see any of the openICPSR pages for my detail as they detail changes I've made in the data such as decisions on what level to aggregate to and mistakes that I made and later found and fixed.

Chapter 3

Offenses Known and Clearances by Arrest

The Offenses Known and Clearances by Arrest dataset - often called Return A, "Offenses Known" or, less commonly, OKCA - is the oldest and most commonly used dataset and measures crimes reported to the police. For this reason it is used as the main measure of crime in the United States and I tend to call it the "crimes dataset." This data includes the monthly number of crimes reported to the police or otherwise known to the police (e.g. discovered while on patrol) for a small selection of crimes, as well as the number of crimes cleared by arrest or by "exceptional means" (a relatively flawed and manipulable measure of whether the case is "solved"). It also covers the number reported but found by police to have not occurred. Since this data has monthly agency-level crime information it is often used to measure crime trends between police agencies and over time. The data uses something called a Hierarchy Rule which means that in incidents with multiple crimes, only the most serious is recorded - though in practice this affects only a small percent of cases, and primarily affects property crimes.

3.1 A brief history of the data

3.1.1 Changes in definitions

3.2 What does the data look like?

3.3 What variables are in the data?

3.3.1 Key variables

3.3.2 Known issues with the data

3.4 Final thoughts

3.5 Hierarchy Rule

This dataset uses what is called the Hierarchy Rule where only the most serious crime in an incident is reported (except for motor vehicle theft, which is always included). For example if there is an incident where the victim is robbed and then murdered, only the murder is counted as it is considered more serious than the robbery.

How much does this affect our data in practice? Actually relatively little. Though the Hierarchy Rule does mean this data is an under-count, data from other sources indicate that it isn't much of an under count. The FBI's other data set, the National Incident-Based Reporting System (NIBRS) contains every crime that occurs in an incident (i.e. it doesn't use the Hierarchy Rule). Using this we can measure how many crimes the Hierarchy Rule excludes (Most major cities do not report to NIBRS so what we find in NIBRS may not apply to them). In over 90% of incidents, only one crime is committed. Additionally, when people talk about "crime" they usually mean murder which, while incomplete to discuss crime, means the UCR data here is accurate on that measure.

3.6 Which crimes are included?

If you look back at the output when we ran `names(offenses_known_yearly_1960_2017)` you'll see that it produced five broad categories of columns. The first was

information about the agency including population and geographic info, then came four columns with the same values except starting with "actual", "tot_clr", "clr_18", and "unfound". Following these starting values were 30 crime categories. We'll discuss what each of those starting values mean in a bit, let's first talk about which crimes are included and what that means for research.

3.6.1 Index Crimes

The Offenses Known and Clearances by Arrest data set contains information on the number of "Index Crimes" (sometimes called Part I crimes) reported to each agency. These index crimes are a collection of eight crimes that, for historical reasons based largely by perceived importance in the 1920's when the UCR program was first developed, are used as the primary measure of crime today. Other data sets in the UCR, such as the Arrests by Age, Sex, and Race data and the Hate Crime data have more crimes reported.

The crimes are, in order by the Hierarchy Rule -

1. Homicide
    - Murder and non-negligent manslaughter
    - Manslaughter by negligence
2. Rape
    - Rape
    - Attempted rape
3. Robbery
    - With a firearm
    - With a knife of cutting instrument

- With a dangerous weapon not otherwise specified
- Unarmed - using hands, fists, feet, etc.

4. Aggravated Assault (assault with a weapon or causing serious bodily injury)
    - With a firearm
    - With a knife of cutting instrument
    - With a dangerous weapon not otherwise specified
    - Unarmed - using hands, fists, feet, etc.
5. Burglary
    - With forcible entry
    - Without forcible entry
    - Attempted burglary with forcible entry
6. Theft (other than of a motor vehicle)
7. Motor Vehicle Theft
    - Cars
    - Trucks and buses
    - Other vehicles
8. Arson
9. Simple Assault

For a full definition of each of the index crimes see the FBI's Offense Definitions page here.

Arson is considered an index crime but is not reported in this data - you need to use the separate Arson data set of the UCR to get access to arson counts. The ninth crime on that list, simple assault, is not considered an index crime but is nevertheless included in this data.

Each of the crimes in the list above, and their subcategories, are included in the UCR data. In most reports, however, you'll see them reported as the total number of index crimes, summing up categories 1-7 and reporting that as "crime." These index crimes are often divided into violent index crimes - murder, rape, robbery, and aggravated assault - and property index crimes - burglary, theft, motor vehicle theft.

3.6.2 The problem with using index crimes

The biggest problem with index crimes is that it is simply the sum of 8 (or 7 since arson data usually isn't available) crimes. Index crimes have a huge range in their seriousness - it includes both murder and theft. This is clearly wrong as 100 murders is more serious than 100 thefts. This is especially a problem as less serious crimes (theft mostly) are far more common than more serious crimes (in 2017 there were 1.25 million violent index crimes in the United States. That same year had 5.5 million thefts.). So index crimes under-count the seriousness of crimes. Looking at total index crimes is, in effect, mostly just looking at theft.

This is especially a problem because it hides trends in violent crimes. San Francisco, as an example, has had a huge increase in index crimes in the last several years. When looking closer, that increase is driven almost entirely by the near doubling of theft since 2011. During the same years, violent crime has stayed fairly steady. So the city isn't getting more dangerous but it appears like it is due to just looking at total index crimes.

Many researchers divide index crimes into violent and nonviolent categories, which helps but is still not entirely sufficient. Take Chicago as an example. It is a city infamous for its large number of murders. But as a fraction of index crimes, Chicago has a rounding error worth of murders. Their 653 murders in 2017 is only 0.5% of total index crimes. For violent index crimes,

murder makes up 2.2%. What this means is that changes in murder are very difficult to detect. If Chicago had no murders this year, but a less serious crime (such as theft) increased slightly, we couldn't tell from looking at the number of index crimes.

3.6.3 Rape definition change

The FBI changed the definition of rape for UCR data starting in 2013 to a broader definition than the older definition, which is commonly called the "legacy definition" or "legacy" or "historical" rape. The legacy definition is "the carnal knowledge of a female **forcibly** and against her will" (emphasis added). This means that only rape is only included in UCR data when it is a female (or any age, there is no differentiation for child victims) forcibly vaginally penetrated by a penis. This is a narrow definition and excludes a number of sexual acts that people may consider rape such as forced oral or sex, and cases with a male victim.

The new (and current) definition "penetration, no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim." Starting in 2013, rape has a new, broader definition in the UCR to include oral and anal penetration (by a body part or object) and to allow men to be victims. The new definition is: "Penetration, no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim." The previous definition included only forcible intercourse against a woman. This definition is far broader and is effectively any non-consensual sexual act. It also includes male victims though the data does not differentiate between male or female (or any other gender) victims.

Both the current and legacy definitions exclude statutory rape and incest other than forcible incest. They both also include lack of consent as cases where the victim cannot give consent, such as if they are too young or are mentally or physically incapacitated - they specifically give the example of being temporarily incapacitated through drugs or alcohol.

As this revised definition is broader than the original one post-2013, rape data is not comparable to pre-2013 data. 2013, however, is simply the year that the FBI changed the definition which means that agencies should have

changed their reporting to the new definition. As might not be too surprising, not all agencies followed this requirement. We'll look at three examples to show when there is clear evidence that the agency did change their definition in 2013, when it's clear they did so a year later, and when it's unclear exactly when they made the change.

3.7 Actual offenses, clearances, and unfounded offenses

For each crime we have four different categories indicating the number of crimes actually committed, the number cleared, and the number determined to not have occurred.

3.7.1 Actual

This is the number of offenses that occurred, simply a count of the number of crimes that month. For example if 10 people are murdered in a city the number of "actual murders" would be 10.

3.7.2 Total Cleared

A crime is cleared when an offender is arrested or when the case is considered cleared by exceptional means. When a single offender for a crime is arrested, that crime is considered cleared. If multiple people committed a crime, only a single person must be arrested for it to be cleared, and as the UCR data is at the offense level, making multiple arrests for an incident only counts as one incident cleared. So if 10 people committed a murder and all 10 were arrested, it would report one murder cleared not 10. If only one of these people are arrested it would still report one murder cleared - the UCR does not even say how many people commit a crime.

A crime is considered exceptionally cleared if the police can identify the offender, have enough evidence to arrest the offender, know where the offender is, but is unable to arrest them. Some examples of this are the death of the offender or when the victim refuses to cooperate in the case.

Unfortunately, this data does not differentiate between clearances by arrest or by exceptional means. For a comprehensive report on how this variable can be exploited to exaggerate clearance rates, see this report by ProPublica on exceptional clearances with rape cases.

3.7.3 Cleared Where All Offenders Are Under 18

This variable is very similar to Total Cleared except is only for offenses in which **every** offender is younger than age 18.

3.7.4 Unfounded

An unfounded crime is one in which a police investigation has determined that the reported crime did not actually happen. For example if the police are called to a possible burglary but later find out that a burglary did not occur, they would put it down as 1 unfounded burglary. This is based on police investigation rather than the decision of any other party such as a coroner, judge, jury, or prosecutor.

3.8 Number of months reported

UCR data is reported monthly though even agencies that decide to report their data may not do so every month. As we don't want to compare an agency which reports 12 months to one that reports fewer, the variable *number_of_months_reported* is way keep only agencies that report 12 months, or deal with those that report fewer.

From our `table()` output it seems that when agencies do report, they tend to do so for all 12 months of the year. However, this variable is seriously flawed, and its name is quite misleading. In reality this variable is actually just whichever the last month reported was. If an agency reported every month of the year, meaning December is the last month, they would have a value of 12. If the agency **only** reported in December, they would also have a value of 12. While there are ways in the monthly data to measure actual number of months reported, these ways are also flawed. So be cautious about this data and particularly the value of this variable.

Chapter 4

Arrests by Age, Sex, and Race

The Arrests by Age, Sex, and Race dataset - often called ASR or the "arrests data" - includes the monthly number of arrests for a variety of crimes and, unlike the crime data, breaks down this data by age and gender. This data includes a broader number of crime categories than the crime dataset (the Offenses Known and Clearances by Arrest data) though is less detailed on violent crimes since it does not breakdown aggravated assault or robberies by weapon type as the Offenses Known data does. For each crime it says the number of arrests for each gender-age group with younger ages (15-24) showing the arrestee's age to the year (e.g. age 16) and other ages grouping years together (e.g. age 25-29, 30-34, "under 10"). It also breaks down arrests by race-age by including the number of arrestees of each race (American Indian, Asian, Black, and White) are the only included races) and if the arrestee is a juvenile (<18 years old) or an adult. The data does technically include a breakdown by ethnicity-age (e.g. juvenile-Hispanic, juvenile-non-Hispanic) but almost no agencies report this data (for many years zero agencies report ethnicity) so in practice the data does not include ethnicity. As the data includes counts of arrestees, people who are arrested multiple times are included in the data multiple times - it is not a measure of unique arrestees.

4.1 A brief history of the data

4.1.1 Changes in definitions

4.2 What does the data look like?

4.2.1 Raw data

4.3 What variables are in the data?

4.3.1 Key variables

4.4 Known issues with the data

4.5 Final thoughts

Chapter 5

Law Enforcement Officers Killed and Assaulted (LEOKA)

The Law Enforcement Officers Killed and Assaulted data, often called just by its acronym LEOKA, has two main purposes.[1] First, it provides counts of employees employed by each agency - broken down by if they are civilian employees or sworn officers, and also broken down by gender. And second, it measures how many officers were assaulted or killed (including officers who die accidentally such as in a car crash) in a given month. The assault data is also broken down into shift type and type (e.g. alone, with a partner, on foot, in a car, etc.), the offender's weapon, and type of call they are responding to (e.g. robbery, disturbance, traffic stop). The killed data simply says how many officers are killed feloniously (i.e. murdered) or died accidentally (e.g. car crash) in a given month. The employee information is at the year-level so you know, for example, how many male police officers were employed in a given year at an agency, but don't know any more than that such as how many officers were on patrol, were detectives, were in special units, etc. This dataset is commonly used as a measure of police employees and is a generally reliable - though imperfect as we'll see - measure of how many police are employed by a police agency. The second part of this data, measuring assaults and deaths, is more flawed with missing data issues and data error issues (e.g. more officers killed than employed in an agency).

[1] This data is also sometimes called the "Police Employees" dataset.

5.1 Common uses of this data

This data, as well as the privately-run site Officer Down Memorial Page which covers law enforcement officers who have died, has also been used lately in the context of police using force against people out of fear of being harmed by that person. The discussion revolves around whether police are actually in high danger of being harmed by comparing the rate at which officers die to that of other professions. In general they find that police officers are among the most likely profession to die but are not at the top of this measure.

5.2 A brief history of the data

5.2.1 Changes in definitions

5.3 What does the data look like?

5.4 What variables are in the data?

5.4.1 Key variables

5.5 Known issues with the data

5.6 Final thoughts

Chapter 6

Supplementary Homicide Reports (SHR)

The Supplementary Homicide Reports dataset - often abbreviated to SHR - is the most detailed of the UCR datasets and provides information about the circumstances and participants (victim and offender demographics and relationship status) for homicides.[1] For each homicide incident it tells you the age, gender, race, and ethnicity of each victim and offender as well as the relationship between the first victim and each of the offenders (but not the other victims in cases where there are multiple victims). It also tells you the weapon used by each offender and the circumstance of the killing, such as a "lovers triangle" or a gang-related murder. As with other UCR data, it also tells you the agency it occurred in and the month and year when the crime happened.

While highly detailed compared to other UCR data, there are a number of limitations for this data.

Since this data is voluntary to

This
If this "most detailed" dataset sounds disappointing - and it is! -

[1] If you're familiar with the National Incident-Based Reporting System (NIBRS) data that is replacing UCR, this dataset is the closest UCR data to it, though it is still less detailed than NIBRS data.

6.1 A brief history of the data

The data is available from the FBI starting in 1975 though, unlike all later years, this year only has information on a single victim and a single offender. For this reason I only release data starting in 1976 where up to 11 victims and 11 offenders are included. This data has been released every year since and the most recent year available is 2019.

6.1.1 Changes in definitions

6.2 What does the data look like?

6.2.1 Raw data

6.3 What variables are in the data?

6.3.1 Key variables

6.4 Known issues with the data

6.5 Final thoughts

Chapter 7

Hate Crime Data

This dataset covers crimes that are reported to the police and judged by the police to be motivated by hate More specifically, they are, first, crimes which were, second, motivated - at least in part - by bias towards a certain person or group of people because of characteristics about them such as race, sexual orientation, or religion. The first part is key, they must be crimes - and really must be the selection of crimes that the FBI collects for this dataset. Biased actions that don't meet the standard of a crime, or are of a crime not included in this data, are not considered hate crimes. For example, if someone yells at a Black person and uses racial slurs against them, it is clearly a racist action. For it to be included in this data, however, it would have to extend to a threat since "intimidation" is a crime included in this data but lesser actions such as simply insulting someone is not included. For the second part, the bias motivation, it must be against a group that the FBI includes in this data. When this data collection began crimes against transgender people were not counted so if a transgender person was assaulted or killed because they were transgender, this is not a hate crime recorded in the data.[1]

So this data is really a narrower measure of hate crimes than it might seem. In practice it is (some) crimes motivated by (some) kinds of hate that are reported to the police. It is also the most under-reported UCR dataset with most agencies not reporting any hate crimes to the FBI. This leads to huge gaps in the data with some states having zero agencies report crime, agencies

[1]The first year where transgender as a group was a considered a bias motivation was in 2014.

reporting some bias motivations but not others, agencies reporting some years but not others. While these problems exist for all of the UCR datasets, it is most severe in this data. This problem is exacerbated by hate crimes being rare even in agencies that report them - with such rare events, even minor changes in which agencies report or which types of offenses they include can have large effects.

My main takeaway for this data is that it is inappropriate to use it to study hate crimes. At most it can be used to look at within-city within-bias-motivation trends, while keeping in mind that even this narrow subset of data is limited by under-reporting by victims and potential changes in police practices of reporting such as how many months of data they report per year.

7.1 A brief history of the data

7.1.1 Changes in definitions

7.2 What does the data look like?

7.2.1 Raw data

7.3 What variables are in the data?

7.3.1 Key variables

7.4 Known issues with the data

7.5 Final thoughts

Chapter 8

Property Stolen and Recovered (Supplement to Return A)

The Property Stolen and Recovered data - sometimes called the Supplement to Return A (Return A being another name for the Offenses Known and Clearances by Arrest dataset, the "crime" dataset) - provides monthly information about property-related offenses (theft, motor vehicle theft, robbery, and burglary), including the location of the offense (in broad categories like "gas station" or "residence), what was stolen (e.g. clothing, livestock, firearms), and how much the stolen items were worth.[1] The"recovered" part of this dataset covers the type and value of property recovered so you can use this, along with the type and value of property stolen, to determine what percent and type of items the police managed to recover. Like other UCR datasets this is at the agency-month level so you can, for example, learn how often burglaries occur at the victim's home during the day, and if that rate changes over the year or differs across agencies. The data, however, provides no information about the offender or the victim (other than if the victim was an individual or a commercial business [based on the location of the incident - "bank", "gas station", etc]). The value of the property stolen is primarily based on the victim's estimate of how much the item is worth (items that are decreased in value once used - such as cars - are supposed to be valued at the current market rate, but the data provides no indication of when it

[1] It also includes the value of items stolen during rapes and murders, if anything was stolen.

uses the current market rate or the victim's estimate) so it should be used as a very rough estimate of value.

8.1 A brief history of the data

8.1.1 Changes in definitions

8.2 What does the data look like?

8.2.1 Raw data

8.3 What variables are in the data?

8.3.1 Key variables

8.4 Known issues with the data

8.5 Final thoughts

Chapter 9

Arson

The arson dataset provides monthly counts at the police agency-level for arsons that occur, and includes a breakdown of arsons by the type of arson (e.g. arson of a person's home, arson of a vehicle, arson of a community/public building) and the estimated value of the damage caused by the arson. This data includes all arsons reported to the police or otherwise known to the police (e.g. discovered while on patrol) and also has a count of arsons that lead to an arrest (of at least one person who committed the arson) and reports that turned out to not be arsons (such as if an investigation found that the fire was started accidentally). For each type of arson it includes the number of arsons where the structure was uninhabited or otherwise not in use, so you can estimate the percent of arsons of buildings which had the potential to harm people. This measure is for structures where people normally did not inhabit the structure - such as a vacant building where no one lives. A home where no one is home *at the time of the arson* does not count as an uninhabited building. In cases where the arson led to a death, that death would be recorded as a murder on the Offenses Known and Clearances by Arrest dataset - but not indicated anywhere on this dataset. If an individual who responds to the arson dies because of it, such as a police officer or a firefighter, this is not considered a homicide (though the officer death is still included in the Law Enforcement Officers Killed and Assaulted data) unless the arsonists intended to cause their deaths. Even though the UCR uses the Hierarchy Rule, where only the most serious offense that occurs is recorded, all arsons are reported - arson is except fro the Hierarchy Rule.

Chapter 10

County-Level Detailed Arrest and Offense Data

10.1 A brief history of the data

10.1.1 Changes in definitions

10.2 What does the data look like?

10.2.1 Raw data

10.2.2 Cleaned data

10.3 What variables are in the data?

10.3.1 Key variables

10.3.2 Known issues with the data

10.4 Final thoughts

Bibliography

Bennice, J. A. and Resick, P. A. (2003). Marital rape: History, research, and practice. *Trauma, Violence, & Abuse*, 4(3):228–246.

Devito, L. (2020). Marijuana arrests on the decline but still outnumber violent crime arrests, according to FBI data. https://www.citybeat.com/news/blog/21145113/marijuana-arrests-on-the-decline-but-still-outnumber-violent-crime-arrests-according-to-fbi-data.

Earlenbaugh, E. (2020). More people were arrested for cannabis last year than for all violent crimes put together, according to FBI data. https://www.forbes.com/sites/emilyearlenbaugh/2020/10/06/more-people-were-arrested-for-cannabis-last-year-than-for-all-violent-crimes-put-together-according-to-fbi-data/?sh=7dd59ad6122f.

Edwards, E., Greytak, E., Madubuonwu, B., Sanchez, T., Beiers, S., Resing, C., Fernandez, P., and Sagiv, G. (2020). A tale of two countries: Racially targeted arrests in the era of marijuana reform. https://www.aclu.org/report/tale-two-countries-racially-targeted-arrests-era-marijuana-reform.

Ingraham, C. (2016). Police arrest more people for marijuana use than for all violent crimes — combined. https://www.washingtonpost.com/news/wonk/wp/2016/10/12/police-arrest-more-people-for-marijuana-use-than-for-all-violent-crimes-combined/.

Kertscher, T. (2019). More arrests in the us for marijuana possession than violent crimes, 2020 hopeful cory booker says.

https://www.politifact.com/factchecks/2019/jul/11/cory-booker/more-arrests-us-marijuana-possession-violent-crime/.

McMahon-Howard, J., Clay-Warner, J., and Renzulli, L. (2009). Criminalizing spousal rape: The diffusion of legal reforms. *Sociological Perspectives*, 52(4):505–531.

Neusteter, S. R. and O'Toole, M. (2019). Every three seconds: Unlocking police data on arrests. https://www.vera.org/publications/arrest-trends-every-three-seconds-landing/arrest-trends-every-three-seconds/overview.

Speri, A. (2019). Police make more than 10 million arrests a year, but that doesn't mean they're solving crimes. https://theintercept.com/2019/01/31/arrests-policing-vera-institute-of-justice/.